Variational Analysis, PDE's and Image analysis: a sampling of details and the big picture
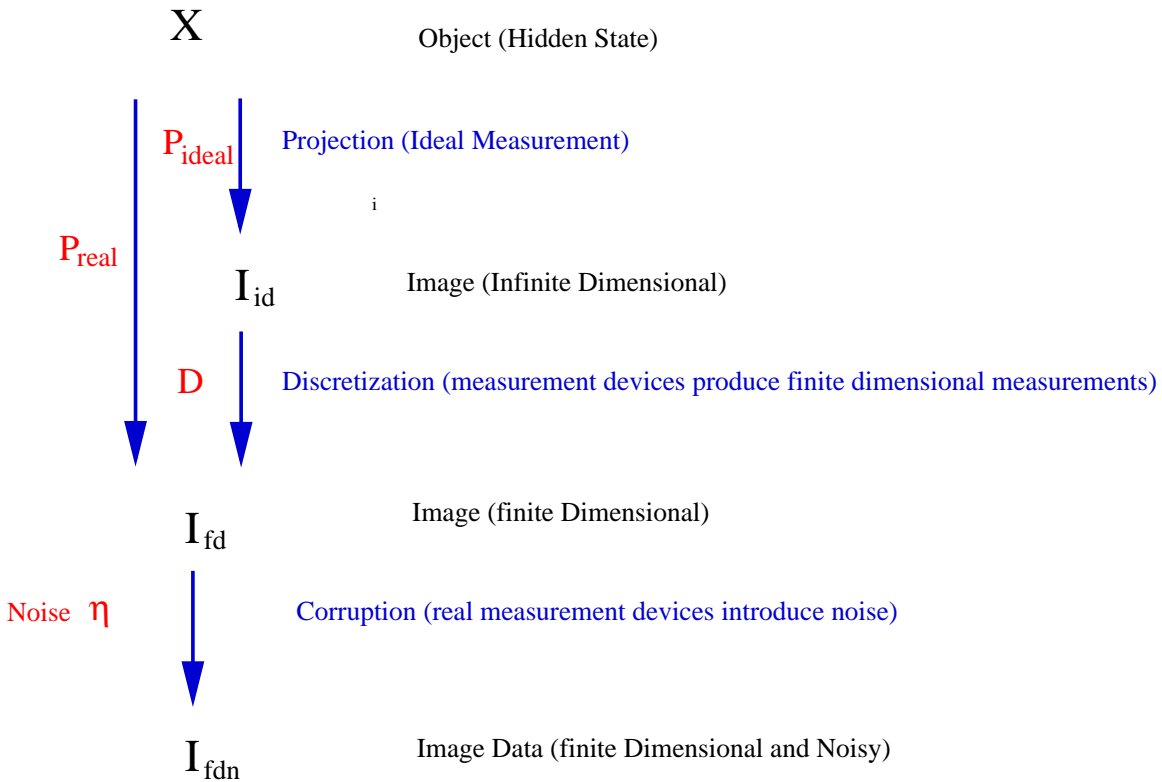
Kevin R. Vixie

Los Alamos National Laboratory

Perspective

When I was a student it was fashionable to give courses called "Elementary Mathematics from a Higher Point of View" ... . But what I needed was a few courses called "Higher Mathematics from the Elementary Point of View".

Joel Franklin

Goal Give a feeling for the way a mathematician thinks about and uses the abstract structures and tools in understanding problems in geometry and analysis applied to the analysis of image data.

View from above: Inverse Problems, Regularization, Data Models and Images

$$\mathrm{X}$$

Object (Hidden State)

P_{ideal} Projection (Ideal Measurement)

P_{real}

i

I_{id} Image (Infinite Dimensional)

D Discretization (measurement devices produce finite dimensional measurements)

I_{fd} Image (finite Dimensional)

Noise η Corruption (real measurement devices introduce noise)

I_{fdn} Image Data (finite Dimensional and Noisy)

We will focus on the noise removal and the reconstruction of hidden information by the use of prior information about the object and/or the infinitely resolved image: We will be interested in the inversion of one of the following three equations.

$$I_{fdn} = P_{real}(x) + \eta$$
$$I_{fdn} = D(I_{id}) + \eta$$
$$I_{fdn} = I_{fd} + \eta$$

Where we start

First we simplify notation:

d is the measured image data (I_{fdn} on the previous page)

u is the object, reconstructed image or denoised image ... (x, I_{id}, I_{fd}, I_{fdn} previously). The space of all such u will be denoted by B and will in general be a Banach space.

P is any measurement or projection operator

With this notation we will be mostly concerned with equations of the type:

$$\min_u F(u) = \min_u ||Pu - d|| + R(u)$$

$R(u)$ **is the regularization term incorporating prior knowledge.**

Note: By differentiating (Gâteaux of Frèchet) F, we can turn a minimization problem into a PDE:

$$\min_u F(u) \Rightarrow u_t = -DF(u)$$

where DF is the derivative of F.

What makes these minimization problems mathematically interesting?

Answer: u lives in an infinite dimensional space

# Banach Spaces, Compactness, Convexity, Coercivity, and all that ...

In this section, I give an idea of ...

- Infinite dimensional $\Rightarrow$ convergence and topologies "bifurcate" – geometry of Banach spaces is highly nontrivial.
- Convexity gives us existence; Strict convexity yields uniqueness.
- Lower semi-continuity plus coercivity plus the weak topology plus reflexivity yields existence of minimizers.
- Derivatives still work in Banach spaces.

Banach Spaces and their Dual Spaces

Banach Space A Banach space B is a vector space with a norm that makes it into a complete metric space.

Dual Space The dual space B^* is simply the space of all bounded linear maps $l : B \to \mathbb{R}$.
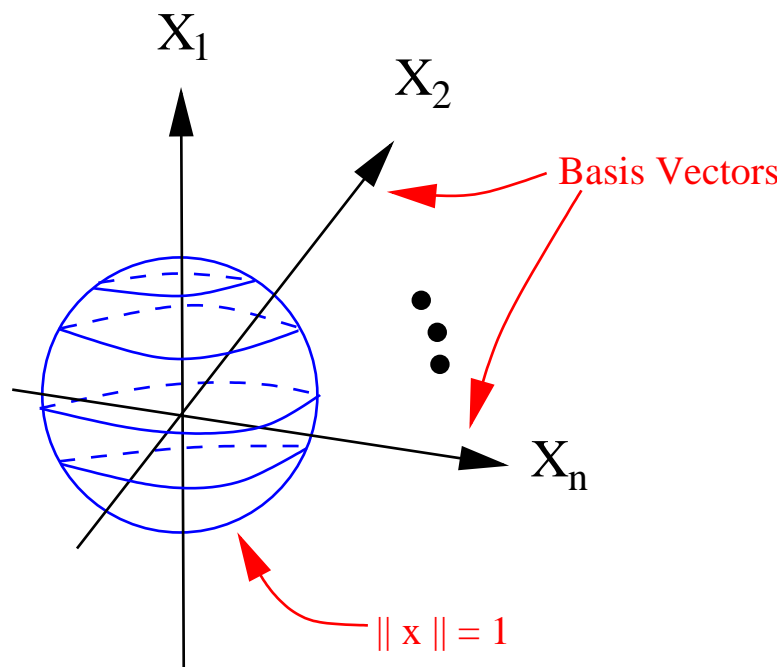

Hilbert spaces are special Banach spaces: Angles are defined, all separable Hilbert spaces are isometrically isomorphic, L^2 and l^2 are both Hilbert spaces, Hilbert spaces are reflexive ...

Compactness

For our purposes, a set is compact if every sequence of points in the set has a convergent subsequence.

In finite dimensional spaces, compact = closed and bounded. In infinite dimensions this is NOT so – the unit ball (in fact any closed ε-ball) is NOT compact.
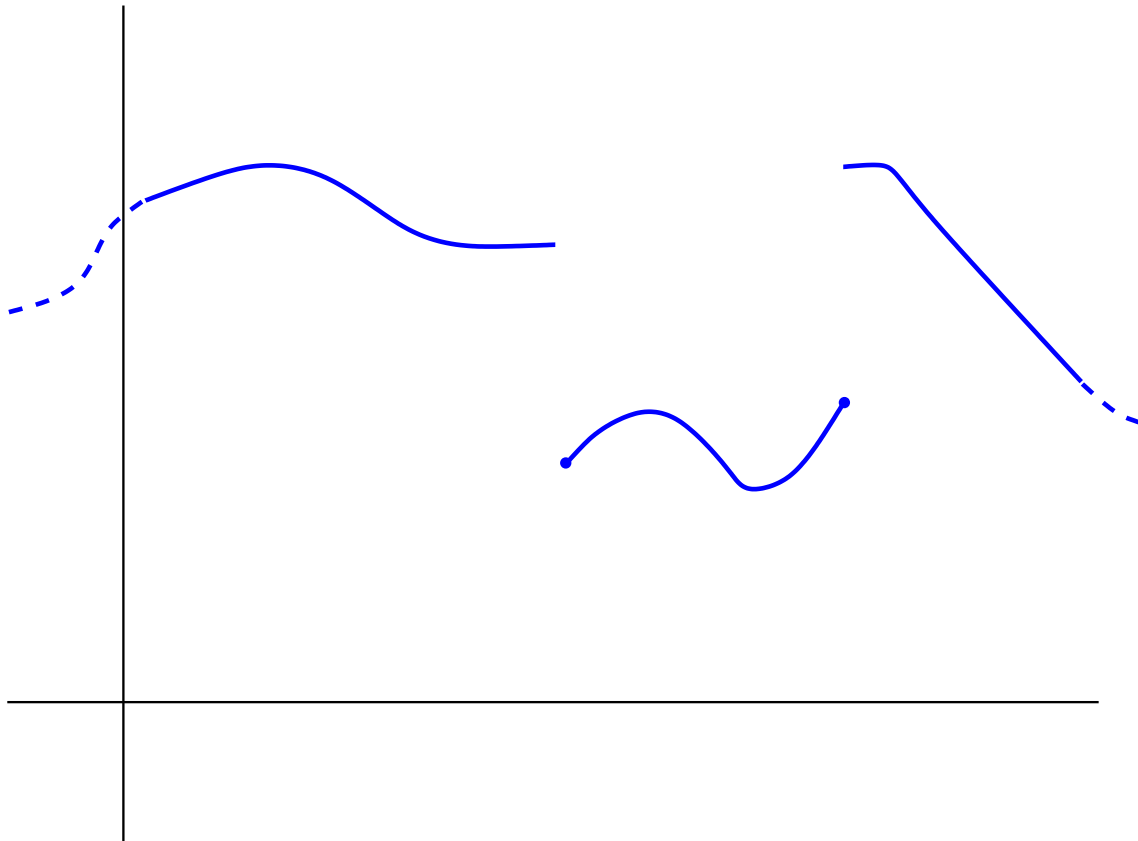
Convexity
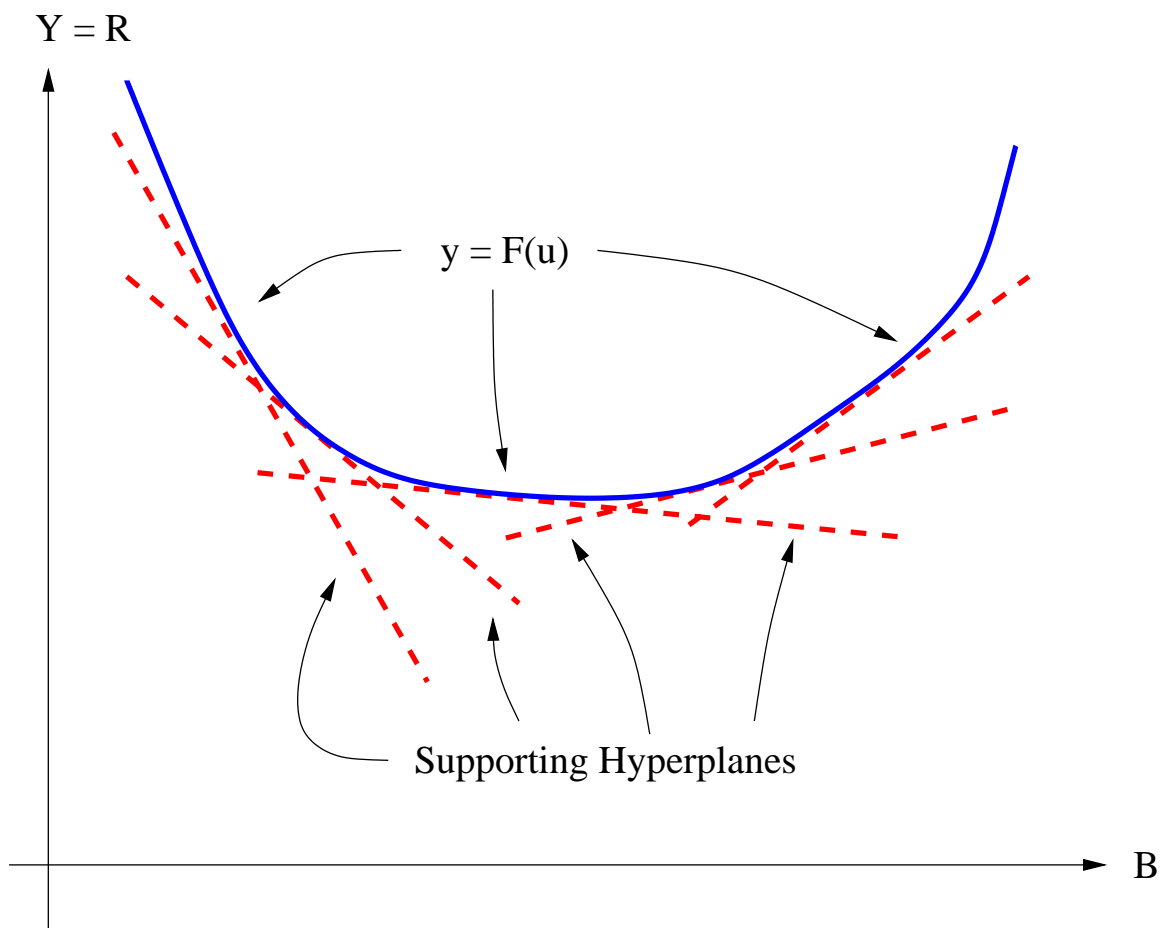
A function is *convex* if for all u_1 and u_2 and every $0 < \alpha < 1$,

$$\alpha f(u_1) + (1 - \alpha)f(u_2) \leq f(\alpha u_1 + (1 - \alpha)u_2).$$

It is said to be *strictly convex* if the above inequality is strict. (The graph of the function is nowhere flat).

Convex functions are to optimization what linear systems are to differential equations.

Lower Semi-Continuity

A function f is *Lower Semi-Continuous* (l.s.c.) if

$$f(u_0) \leq \liminf_{i \to \infty} f(u_i)$$

for all sequences $u_i \to u_0$.

Convergence

- Strong Convergence: $u_i \to u_0$ if $||u_i - u_0|| \to 0$ as $i \to \infty$.
- Weak Convergence: $u_i \underset{w}{\to} u_0$ if $l(u_i) \to l(u_0)$ as $i \to \infty$ for all $l \in B^*$.

The primary practical reason for using other forms of convergence is that though one may lose the nice properties of a metric space, the weak topology recaptures compactness for closed and bounded sets (there are details!).

Example: B, the space of L^2 functions on a measure space (Ω, μ)

$$||u|| = \left(\int_\Omega |u(x)|^2 d\mu(x) \right)^{\frac{1}{2}}$$

Due to the Riesz Representation theorem, we can express the dual space B^* as all linear functionals of the form:

$$l_g(u) = \left(\int_\Omega g(x) u(x) d\mu(x) \right)$$

Strong Convergence

$u_i \to u_0$ in the strong topology if

$$\lim_{i \to \infty} ||u_i - u_0|| = \lim_{i \to \infty} \left(\int_\Omega |u_i(x) - u_0(x)|^2 d\mu(x) \right)^{\frac{1}{2}} = 0$$

Weak Convergence

$u_i \to u_0$ in the weak topology if

$$\lim_{i \to \infty} l_g(u_i) - l_g(u_0) = \lim_{i \to \infty} \int_\Omega g(x)(u_i(x) - u_0(x)) d\mu(x) = 0$$

for all $l_g \in B^*$ (or equivalently, for all $l_g, g \in B$).

Coercivity

A function F mapping elements from a Banach space B to $\mathbb{R}$ is coercive if the level sets are bounded or equivalently if $F(u) \to \infty$ as $||u|| \to \infty$

Differentiability and Subgradients

f' = −0.5

∂f = [−2, 2]
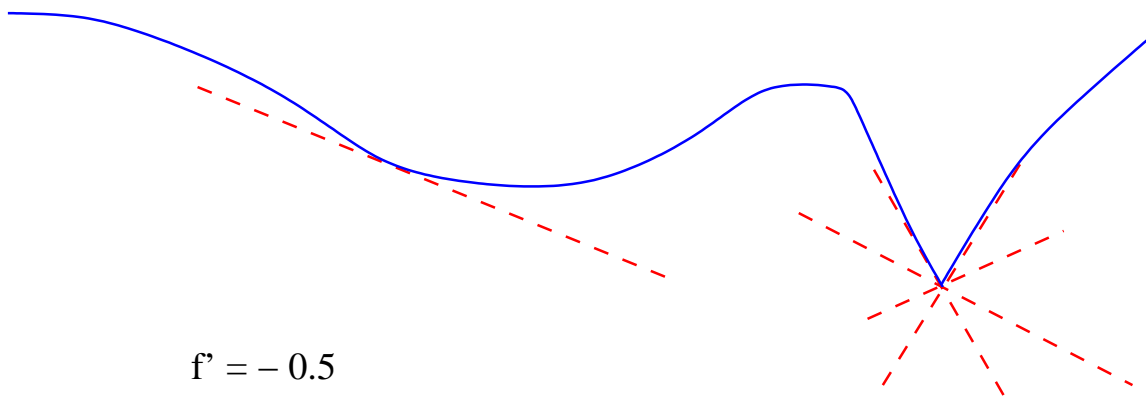
Derivative

Local Subdiferential

Locally Linear Derivatives are local linear approximations: this works in infinite dimensional spaces as easily as in one dimensional spaces. Suppose that:

$$F : \mathbb{R}^3 \to \mathbb{R}^2$$

Then, DF will be a linear operator from $\mathbb{R}^3$ to $\mathbb{R}^2$ and can be represented by a matrix with 3 columns and 2 rows.

Derivatives and the Dual Space If $F : B \to \mathbb{R}$ then DF is in the dual space B^*.

Infinite dimensions is where the zoo itself gets interesting

property	finite dimensional	infinite dimensional
continuity	all linear maps continuous	bounded linear maps continuous
compactness	the unit ball is compact	the unit ball is NOT compact
norms	All norms are equivalent	many different norms on the same space
topologies	Strong $=$ Weak $=$ Weak*	Strong $\neq$ Weak $\neq$ Weak*
sequences	sequences completely characterize topology	need *nets*
linear subspaces	always closed	not necessarily closed

Culmination of the previous concepts: What we need for existence

When can we find a minimizing u^*? i.e. when can we find a u^* such that
$$F(u^*) = \bar{F} \equiv \inf_u F(u).$$
suppose that:

A1 $B = B^{**}$ ($\Rightarrow$ bounded sets are compact under weak convergence)

A2 F is coercive (level sets are bounded), and

A3 F is weakly sequentially lower semi-continuous,

then we may follow the steps:

1 Note that there is a sequence u_i such that $F(u_i) \to \bar{F}$.

2 Since F is coercive, the level sets of F are bounded.

3 Under the weak topology, bounded sets are sequentially compact SO there is a subsequence of u_i call it u_{i_n} and a point u^* such that $u_{i_n} \to_w u^*$ as $n \to \infty$.

4 We use the weak sequential lower semi-continuity to obtain that

$$F(u^*) \leq \liminf_{n \to \infty} F(u_{i_n})$$

5 This implies that $F(u^*) = \bar{F}$.

Consequently, if any particular F satisfies A1, A2, and A3, we automatically have the existence of a minimizer u^*.

Also: Uniqueness

Strict Convexity If F is strictly convex, any minimizer is unique: if not then $u_1^* \neq u_2^*$ and
$F^* \leq F(\lambda u_1^* + (1-\lambda)u_2^*) < \lambda F(u_1^*) + (1-\lambda)F(u_2^*) \leq F^*$ or $F^* < F^*$.

Practical Uniqueness
  1 How big is the bounding ball in which the minimal solutions live?
  2 How big does this bounding ball get as we look for all solutions with F values within ε of F^*?
  3 How big is the set of u satisfying $F_{\hat{d}}(u) \leq F^* + \varepsilon$ for some $\hat{d}$ within δ of the measured data d.

What we need: Stability

Key question How big are the level sets of F at or near the minimal value of F?

If $F(u)$ is convex and the subdifferential at a minimizing u^* contains an ε ball around the origin of the dual space B^* then we know that at least our instability is no worse than $1/\varepsilon$.

Before we move on to examples ...

If you are wishing I had never opened your eyes to all the things that are needed to understand "easy" things like minimization of functions, you might take comfort from:

Everyone knows what a curve is, until he has studied enough mathematics to become confused by the countless exceptions.

Felix Klein

Example: BV Functions and Total Variation Minimization

**Intuition** I give a few insights into what TV minimization is all about

Data I present an example of the use of TV minimization for radiographic data. Inspired by the need to invert sparse radiographic data, we are interested in the following equation:

$$u_{\text{optimal}} \equiv \arg\min_u F(u) = \arg\min_u \left(||Pu - d||^2 + \int_\Omega |\nabla u| \right)$$

We now explain why this is appropriate.

We consider the (1-dimensional) continuous case:

we compute $F(u) \equiv \int_\Omega |\nabla u|^p dx$
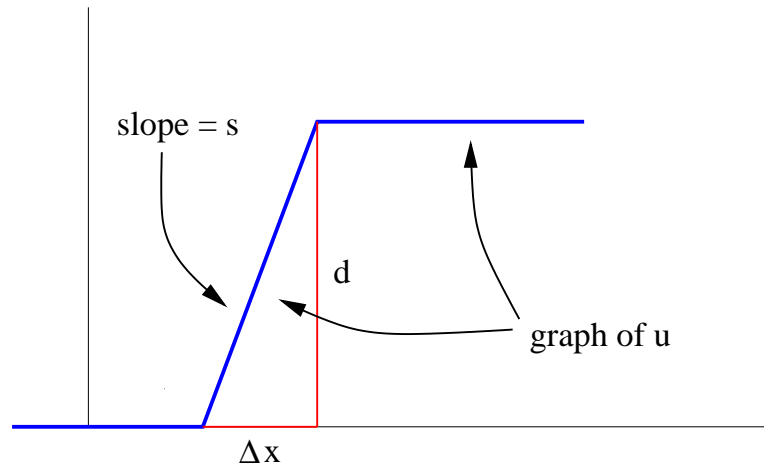
$F(u) = s^p(\Delta x) = \frac{(s\Delta x)^p}{(\Delta x)^{p-1}} = d^p(\Delta x)^{1-p}$

$(p > 1)$ $F(u) \underset{\Delta x \to 0}{\to} \infty$

$(p = 1)$ $F(u) = d$

$(p < 1)$ $F(u) \underset{\Delta x \to 0}{\to} 0$

Moral of the Story:

- For $p > 1$ discontinuities are avoided ... smooth u preferred,
- For $p < 1$ discontinuities cost nothing ... step u preferred,
- BUT for $p = 1$ only the variation or jump magnitude "counts", no bias towards either smooth or step!

Aside: Lagrange Multiplier Picture

Problem Maximize $f(x)$ subject to $g(x) = c$.

Multiplier method Find stationary points of $L(x, \lambda) \equiv f(x) + \lambda g(x)$

Do ... solve $DL(x, \lambda) = Df(x) + \lambda Dg(x) = 0$

A picture:

Gradient of g

Gradient of f

X^*

$f = 2$

$f = 1$

$f = 0$

$g = 0$
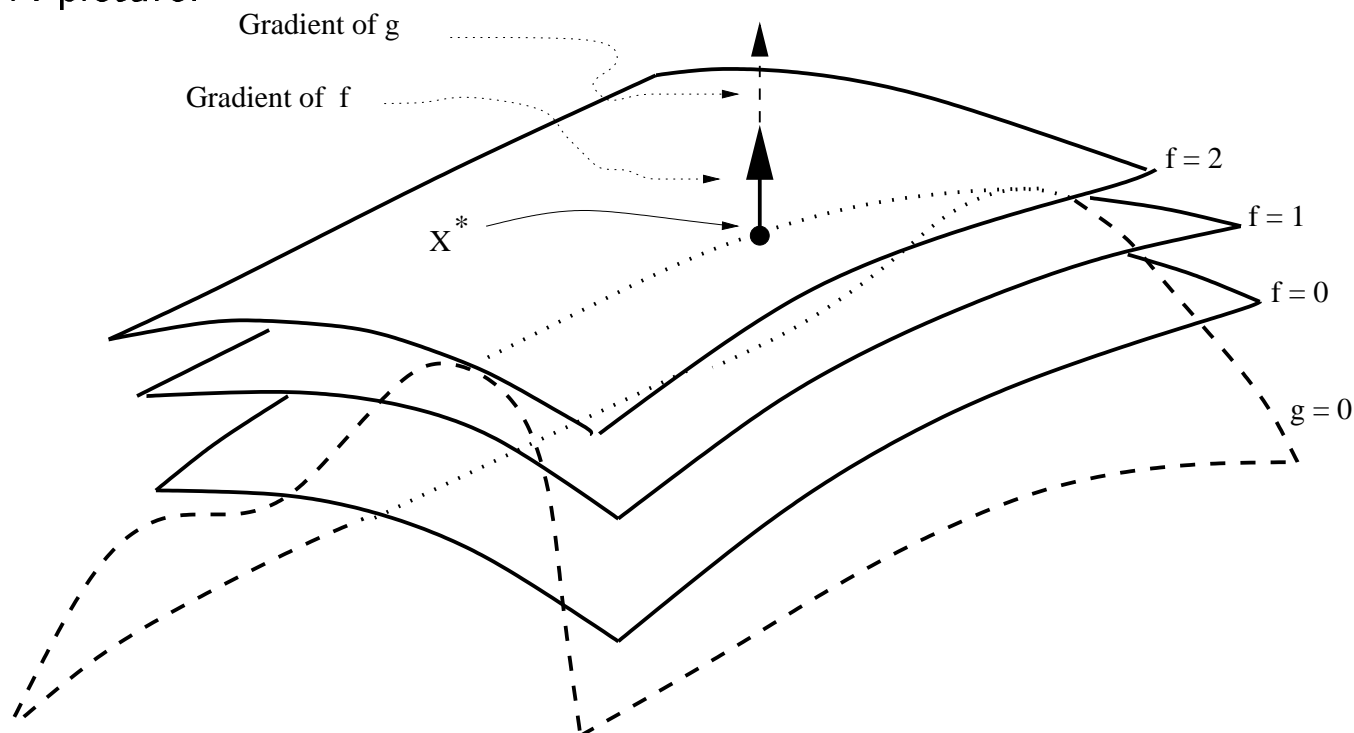
Aside: BV Functions

*W*e can define a space of functions whose norm is based on the measure we introduced to look at edges, $F(f) \equiv \int_\Omega |\nabla f| d\mu$.

The space will be those $f : \mathbb{R}^N \to \mathbb{R}$ such that

$$f \in L^1(\Omega) \text{ and } \int_\Omega |Df| < \infty$$

where we define $\int_\Omega |Df|$ when $f \notin C^1(\Omega)$ by

$$\int_\Omega |Df| \equiv \sup \left\{ \int_\Omega f \text{div}(g) dx : g \in C_c^1(\Omega, \mathbb{R}^N), |g(x)| \leq 1 \ \forall x \in \Omega \right\}$$

In this case we define the BV norm by:

$$||f||_{BV} \equiv ||f||_{L^1} + \int_\Omega |Df|$$

*T*he space of functions f such that $||f||_{BV} < \infty$ together with the BV norm is a Banach space of great utility and versatility.

$TV(f) \equiv \int_\Omega |Df|$ $(= \int_\Omega |\nabla f|$ when $f \in C^1)$ is the *Total Variation* of f.

... a bit more: level sets and feature scale dependence
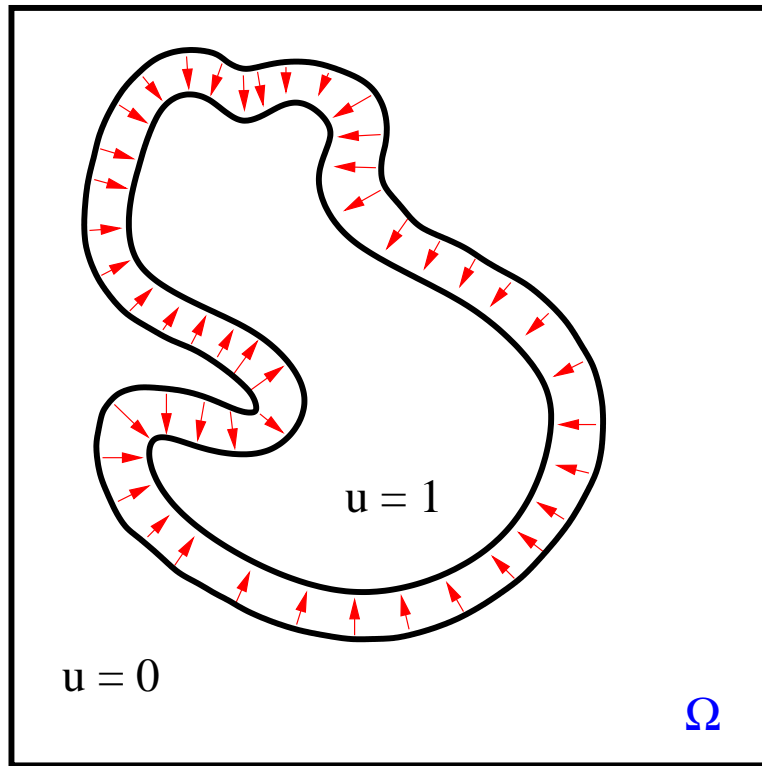
$TV(u)$ can be also computed as an integral over level sets:

$$TV(u) = \int_\Omega |Du| dx dy = \int_\mathbb{R} L(u,r) dr$$

where

$$L(u,r) = \text{length of the boundary of the r-level set.}$$

Total Variation Regularization

*N*ow we consider the image recovery problem and the role that *optimization* and *total variation* play in regularized reconstruction from projections.

A common regularization of the radiographic image reconstruction problem is the use of the L^2 norm to regularize the inverse problem.

$$u_{\text{optimal}} \equiv \arg\min_u F(u) = \arg\min_u \left(||Pu - d||^2 + ||u||_2 \right)$$

where P is the radiographic projection operator and d is the radiographic data. Another regularization is given by the minimization:

$$u_{\text{optimal}} \equiv \arg\min_u F(u) = \arg\min_u \left(||Pu - d||^2 + \int_\Omega |\nabla u|^2 \right)$$

But, as noted above, the $|\nabla u|^2$ is biased against edges, while $|\nabla u|$ is biased neither for nor against edges. This leads us to consideration of:

$$u_{\text{optimal}} \equiv \arg\min_u F(u) = \arg\min_u \left(||Pu - d||^2 + \int_\Omega |\nabla u| \right)$$

So we end up with the following continuous functional and it's discrete counterpart:

$$F(u) = \lambda ||Pu - d||^2 + \int_\Omega |\nabla u|$$

$$\hat{F}(u) = \lambda||Pu - d||^2 + \sum_{i,j}|\nabla_{i,j}u|$$

where $\nabla_{i,j}$ is the discretized gradient. For specific choices of λ we implicitly seek solutions constrained by $||Pu - d||^2 = \sigma_\lambda^2$).

Technically speaking, we need the discretized functional to be strictly convex and coercive for guaranteed existence and uniqueness. It also permits us to prove linear convergence.

A picture: two cylinders

Recalling our pictures of the Lagrange Multipliers and specializing them to our functional $\hat{F}(u)$:

positive span of u = (1,1,1,...,1)

$\| Pu - d\|^2 = \sigma^2$

$N^\perp$

u*

Solution space = u* + N

Level cylinder of TV(u)

N
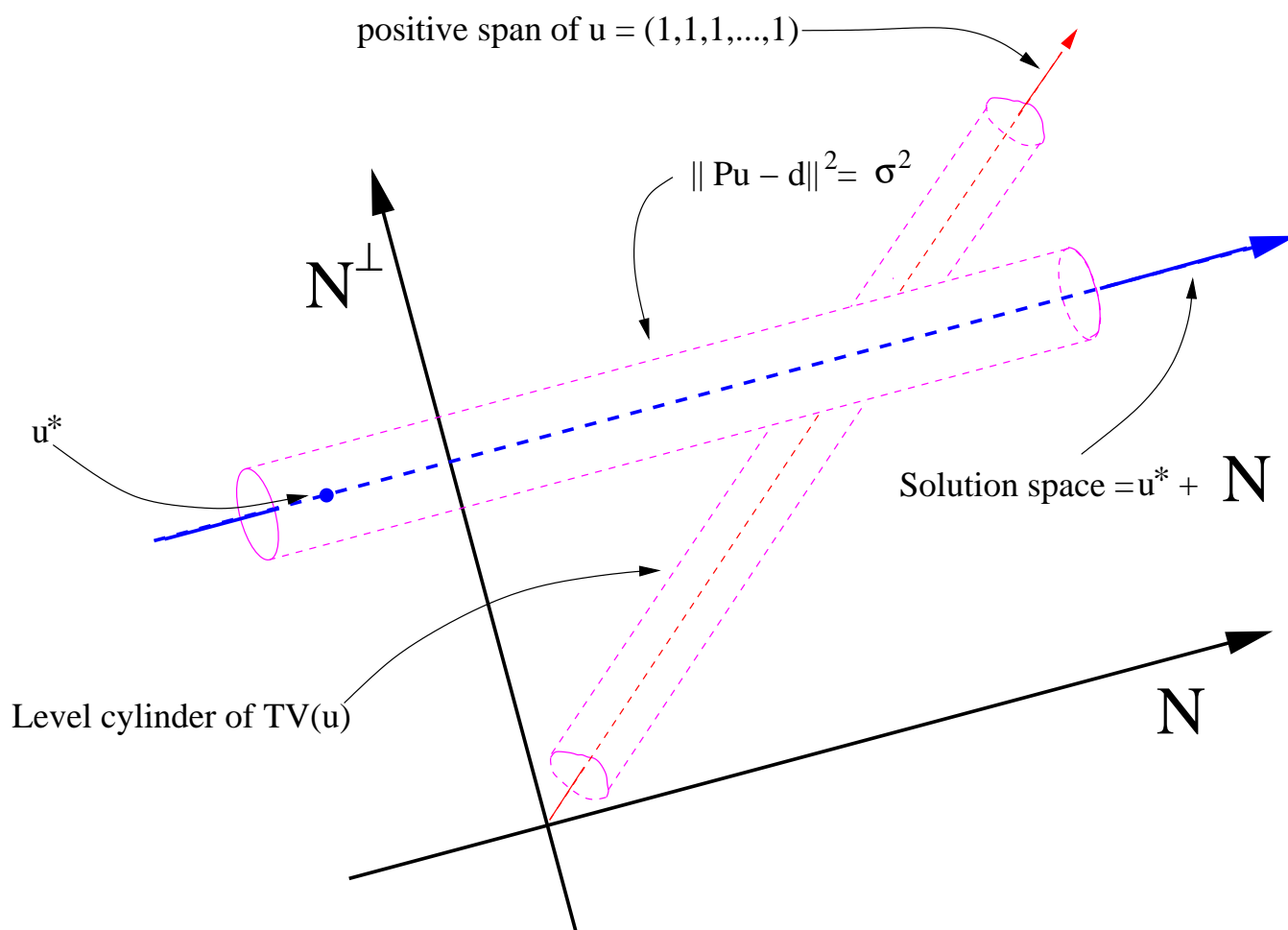
... a closer look ...
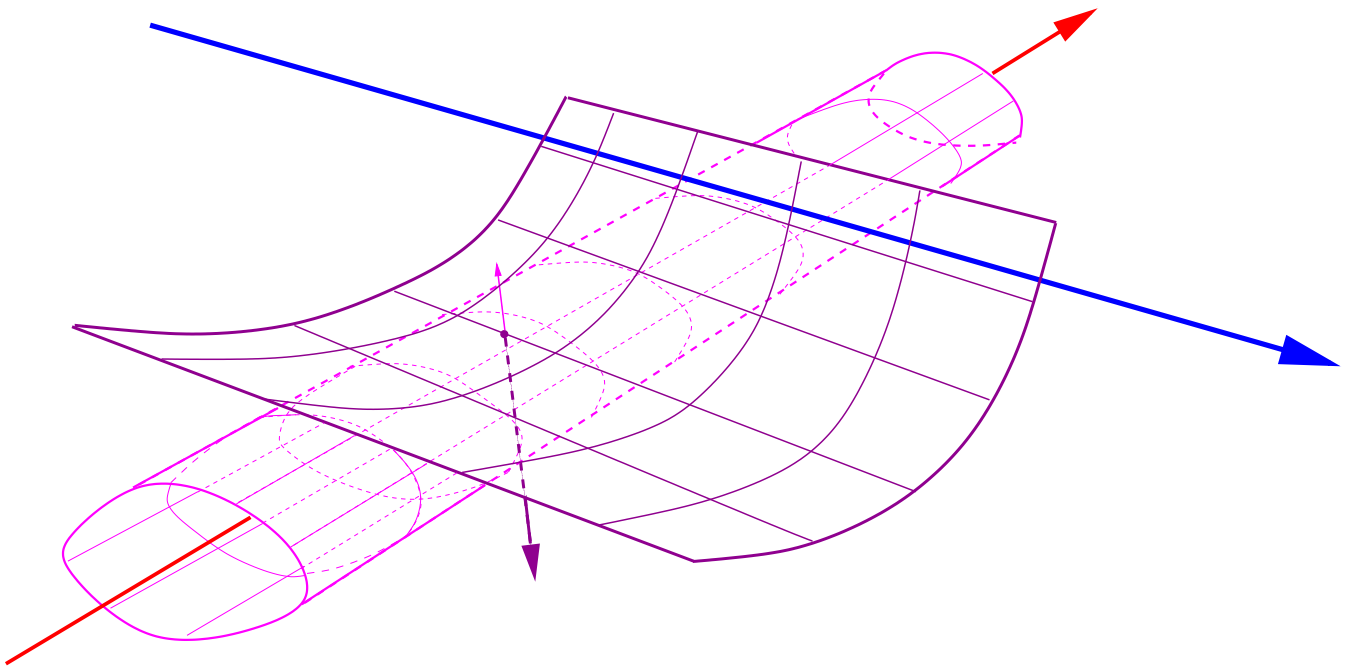
A fixed point method

If we compute the derivative of the functional

$$F(u) = \alpha \sum_{i,j} |\nabla_{i,j} u| + \frac{1}{2} ||Pu - d||^2$$

(where $\alpha = \frac{2}{\lambda}$) and set it to zero, we get

$$\alpha \sum_{i,j} \nabla_{i,j}^T \left(\frac{\nabla_{i,j} u}{|\nabla_{i,j} u|} \right) + P^T Pu - P^T d = 0$$

We turn this into an iterative method that (we can prove) converges to a unique fixed point. The iterative method is given by

$$\alpha \sum_{i,j} \nabla_{i,j}^T \left(\frac{\nabla_{i,j} u_{k+1}}{|\nabla_{i,j} u_k|} \right) + P^T Pu_{k+1} - P^T d = 0$$

At each step we solve for u_{k+1} using a conjugate gradient method. A last modification to remove the singularity in derivative of the TV term is done by noticing that

$$|\nabla u| = \sqrt{|\nabla u|^2} \approx \sqrt{|\nabla u|^2 + \beta}$$

for small β.

Initial Results: BCO4

This data was taken at LANSCE. The work is part of an ongoing project with Tom Asaki, David Strong, and Patrick Campbell.
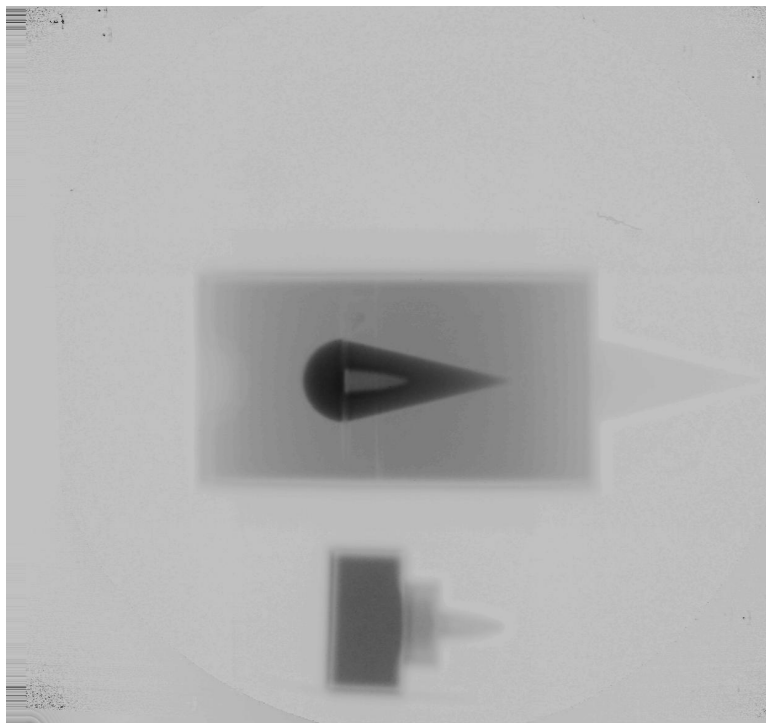
Now we show preliminary results obtained by total variation minimization using the fixed point method. The results were obtained using our own hacks of Pep Mulet's ImageTool, a matlab package.

BCO4 Results: continued
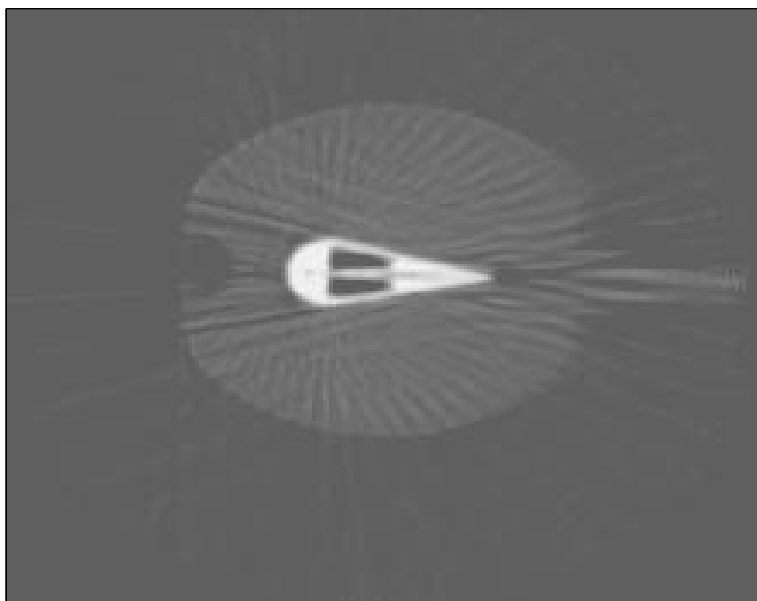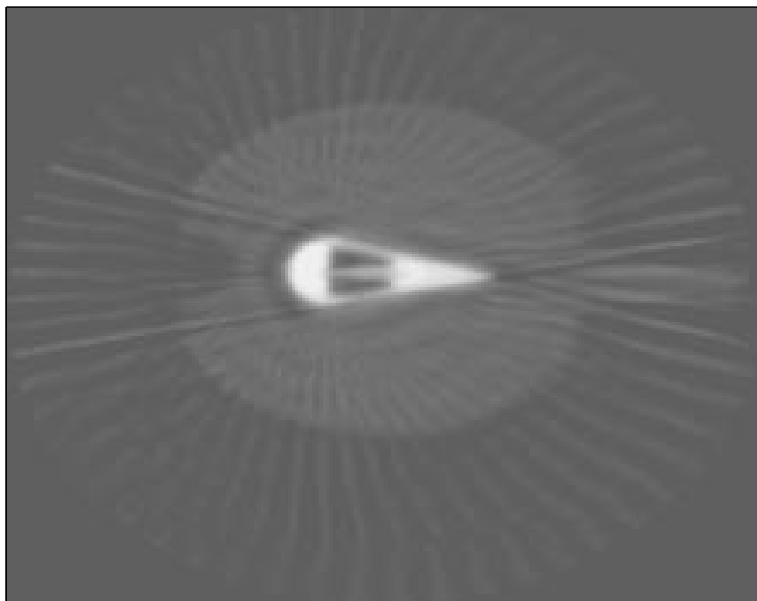
BCO4 Results: continued
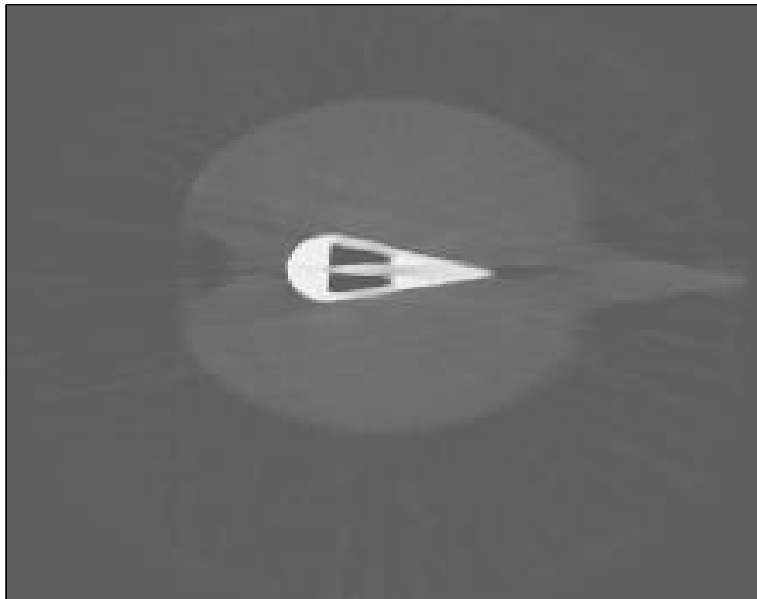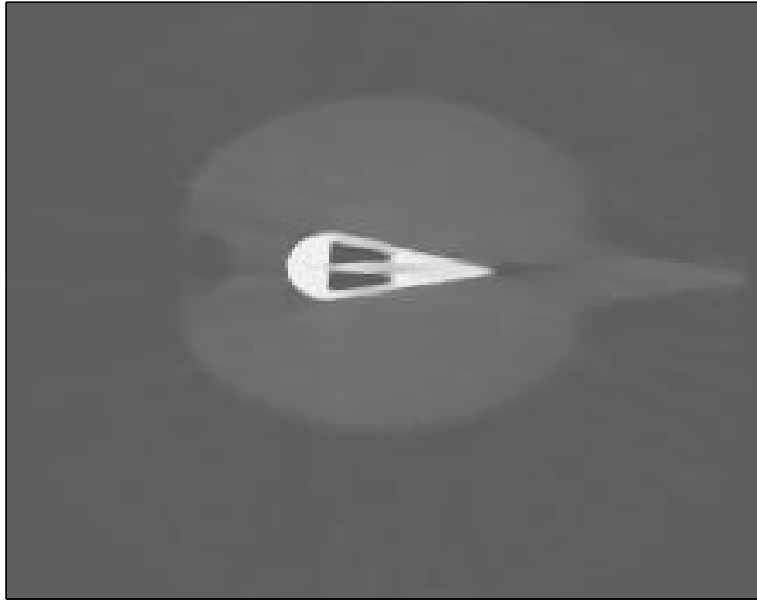
Example: Nonlinear Diffusion and Robust Statistics

One approach to image regularization: Let the measured image d be the initial data (i.e. $u(x,0) = d$) and "run" the heat equation

$$\frac{\partial u}{\partial t} = \nabla \cdot (\nabla u)$$

from $t = 0$ to $t = \tau$.

Results:

1. $u(x,\tau)$ is a smoothed version of the image with length scales suppressed smaller than $\sqrt{\tau}$
2. An undesirable loss of edge information.

Another approach: Perona-Malik (1990) use the above with a modified PDE,

$$\frac{\partial u}{\partial t} = \nabla \cdot (g(||\nabla u||)\nabla u)$$

where g, known as the *stopping function*, is chosen so as to suppress the diffusion at locations with high gradients.

Robust Estimation

Suppose we want to estimate a piecewise constant function from a noisy version of that function.
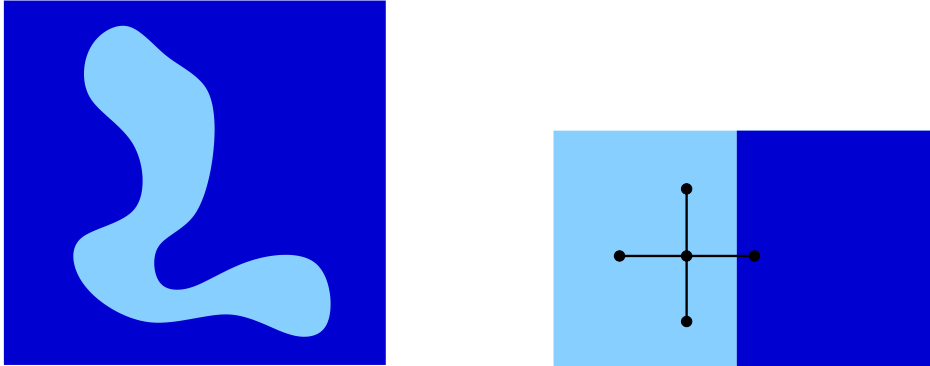
Robust estimation approaches this problem by minimizing

$$\min_u \sum_i \sum_{j \in N_i} \rho(|u_j - u_i|, \sigma)$$

where ρ is a *robust error norm* (think of a negative log likelihood) with intrinsic scale σ.

Idea Choose a norm such that differences above a scale σ are given less weight (or even much less weight).

We would at least expect the case where there is a separation in scales between the noise and the discontinuities to respond well to this strategy.

What is the relation between robust estimation and nonlinear diffusion?

The relation between robust estimation and nonlinear diffusion

Recall that in the case of nonlinear diffusion we are flow the image according to:

$$\frac{\partial u}{\partial t} = \nabla \cdot (g(||\nabla u||)\nabla u)$$

and, noting that the continuous form of the robust estimation problem involves the minimization:

$$\min_u \int_\Omega \rho(||\nabla u||)d\Omega$$

which can be rewritten (via gradient descent as:

$$\frac{\partial u}{\partial t} = \nabla \cdot \left(\rho'(||\nabla u||)\frac{\nabla u}{||\nabla||}\right)$$

so if we let $g(s) = \rho'(s)/s$ we see that *robust estimation is nothing else than nonlinear diffusion and vice versa*.

Example stopping functions

Each set of three plots are g, ψ, and ρ respectively:

- Quadratic (normal)
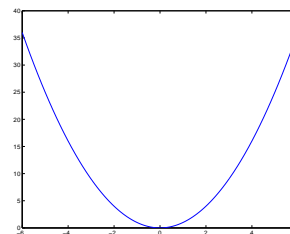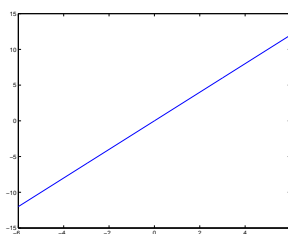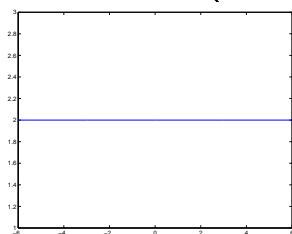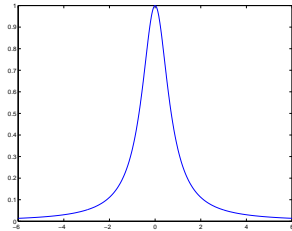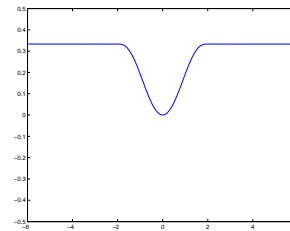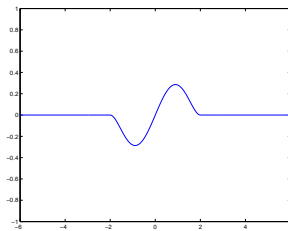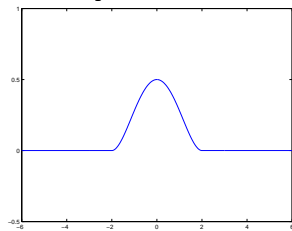
- TV min
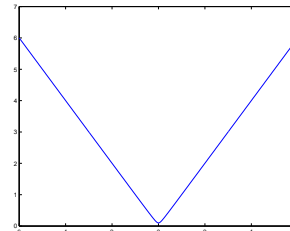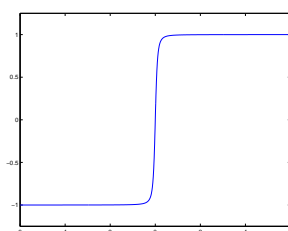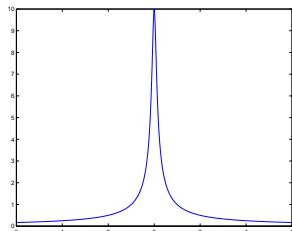
- Lorentzian
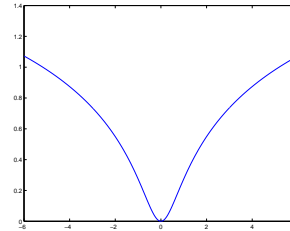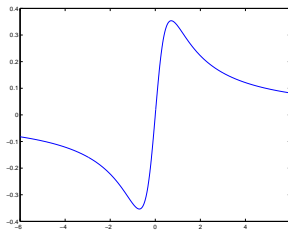
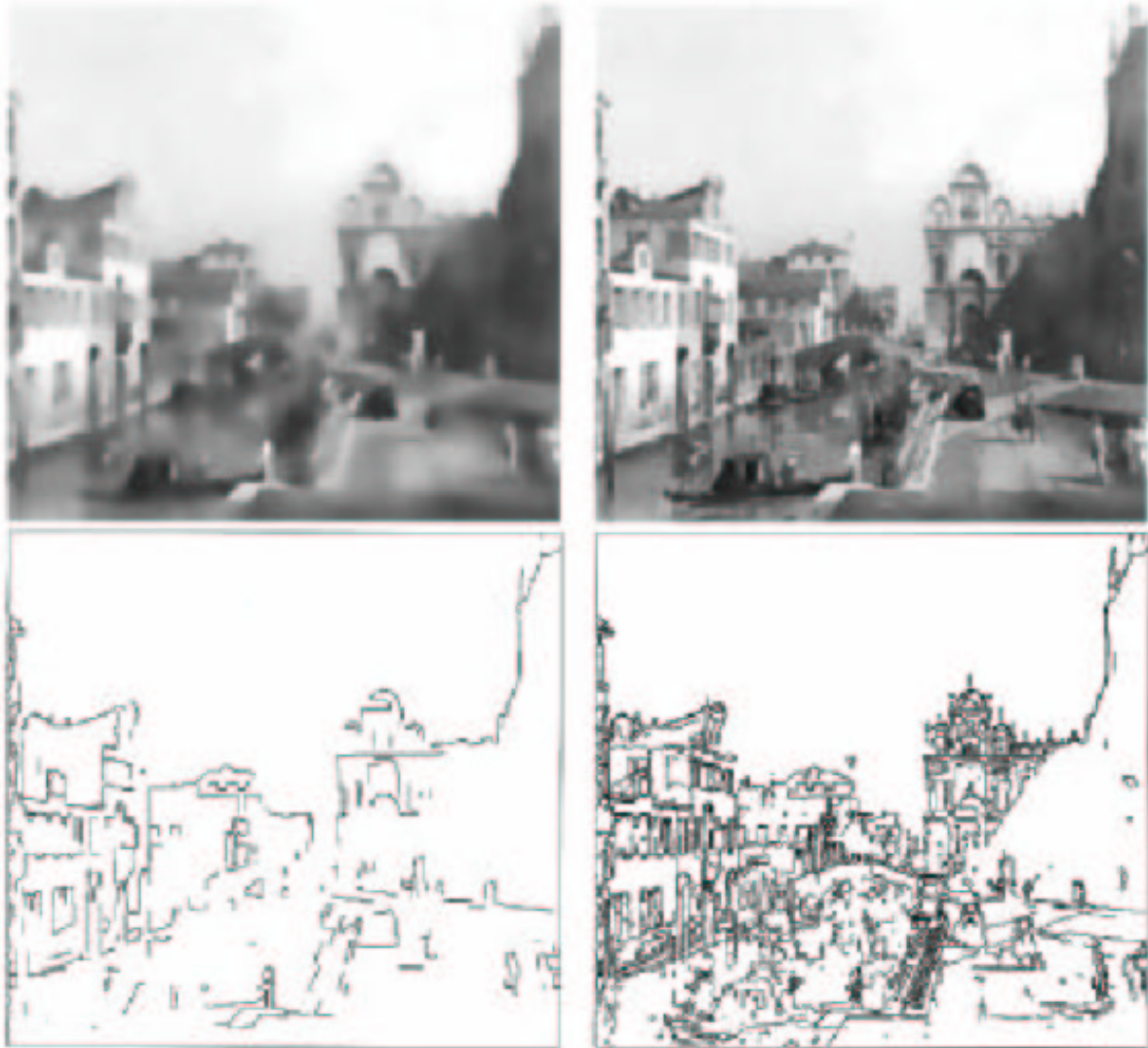- Tukey

Example

(from Black, Saprio, Marimont, and Heeger 1998)

Comparison of the Perona–Malik (Lorentzian) function (left) and the Tukey function (right) after 500 iterations.

Summary

- I have introduced mathematical concepts important for careful consideration of the variational and PDE approach to image analysis.
- I introduced the TV minimization method with examples and then showed that a nonlinear diffusion approach is a generalization of the TV minimization approach. We briefly explored the relation to robust estimation

There is much more, but that is the end of my lecture, after all:

The secret to wearying consists in saying everything

Voltaire

References: Image Analysis and Supporting Mathematics

[1] Papers to be found as CAM reports at UCLA: Chan, Osher, Mulet, Strong, Vese, Strong, ...

[2] Sapiro, "Geometric Partial Differential Equations and Image Analysis"

[3] Aubert and Kornprobst, "Mathematical Problems in Image Processing"

[4] Osher and Fedkiw, "Level Set Methods and Dynamic Implicit Surfaces"

[5] Ambrosio, Fusco, and Pallara, "Functions of Bounded Variation and Free Discontinuity Problems"

[6] Evans and Gariepy, "Measure Theory and Fine Properties of Functions".

[7] Krantz and Parks, "The Geometry of Domains in Space"

[8] Ekeland and Temam, "Convex Analysis and Variational Problems"

[9] Bertsekas, "Nonlinear Programming"

[10] Zeidler, "Nonlinear Functional Analysis and it's Applications III: Variational Methods and optimization"

[11] Vogel, "Computational Methods for Inverse Problems"